Image and Vision Computing 137 (2023) 104750

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

ATOM: Self-supervised human action recognition using atomic motion representation learning

Bruno Degardin [a,b,c,*], Vasco Lopes [a,c], Hugo Proença [a,b]

[a] University of Beira Interior, Portugal
[b] IT - Instituto de Telecomunicações, Portugal
[c] DeepNeuronic, Portugal

ARTICLE INFO

Keywords:
Atomic dynamics
Self-supervised learning
Graph convolutional networks
Human pose
Skeleton-based action recognition
Human behavior understanding

ABSTRACT

Self-supervised learning (SSL) is a promising method for gaining perception and common sense from unlabelled data. Existing approaches to analyzing human body skeletons address the problem similar to SSL models for image and video understanding, but pixel data is far more challenging than coordinates. This paper presents ATOM, an SSL model designed for skeleton-based data analysis. Unlike video-based SSL approaches, ATOM leverages atomic movements within skeleton actions to achieve a more fine-grained representation. The proposed architecture predicts the action order at the frame level, leading to improved perceptions and representations of each action. ATOM outperforms state-of-the-art approaches in two well-known datasets (NTU RGB + D and NTU-120 RGB + D), and its weight transferability enables performance improvements on supervised and semi-supervised tasks, up to 4.4% (3.3% p.p.) and 14.1% (6.3% p.p.), respectively, in Top-1 Accuracy.

1. Introduction

The significant advancements in various image-based tasks and applications have laid a robust groundwork of techniques and expertise to expedite the automation of video comprehension. However, such data-driven approaches [1–4] highly depend on the scale of the learning set, where, naturally, has been an unprecedented increase in data sources [5–9]. Hence, the development of unsupervised and self-supervised learning techniques has become a promising cue to surpass the impracticable labelling process. Despite great success in natural language processing (NLP) [10–13] and image-based [14–16] problems, self-supervised methodologies towards human behaviour are still limited due to their complexity and transferability to supervised models. Current approaches in video understanding have explored pretext tasks, such as contrastive learning [17–19], motion prediction [20], jigsaw puzzle solver [21], video clip order [22–24] and speed prediction [25]. The goal is to establish foundational knowledge and approximate a form of apriori common sense. Variations of these pretext tasks have been proposed in 3D human action recognition [26–29], which often incorporate incremental tasks or modules to improve perception in a self-supervised setting.

However, there are three shortcomings of these skeleton-based methods. (1) Imprecise representations by applying SSL video-based approaches to skeleton data. Since videos are highly complex data, methods similar to [22,25] become suboptimal when applied to skeletons. These models were trained using SSL strategies specifically devised for video data, and thus do not capture the fine grained information of the human pose. (2) The computational and task concept complexity. The popularity of multi-task and meta-learning led to current techniques apply additional pre-text tasks [27,29] for self-supervision in action recognition. However, as previously reported [33] pre-training followed by fine-tuning on single-tasks outperforms multi-task learning due to the increased complexity of a self-supervision paradigm. (3) The difficulty of architecture transferability. Due to increased task complexity, recent methods are introducing supplementary modules [34,27] into the architectures to pre-train (e.g., action prediction and reconstruction). Although effective in resolving unlabelled tasks, the difficulty of replicating and combining these methods across various architectures increases, along with the challenge of transferring weights to supervised tasks due to architectural differences. Following the well-known taxonomy of human behaviour, [35,36], this paper proposes an atomic motion representation learning approach (ATOM) by analyzing the lowest hierarchy level of motion (over an action execution), also known as atomic dynamics, to address the above-mentioned limitations. Our

* Corresponding author at: University of Beira Interior, Portugal.
E-mail address: bruno.degardin@ubi.pt (B. Degardin).

https://doi.org/10.1016/j.imavis.2023.104750
Received 13 December 2022; Received in revised form 11 April 2023; Accepted 20 June 2023
Available online 28 June 2023

methodology leverages the atomic dynamics over human behaviour at the frame level to improve learning and solve spatial and temporal relationships without any label and supplementary architecture modules or multiple tasks. Inspired by self-supervised learning of language representations through sentence order prediction [10,11,13], we split an action into smaller chunks and even frames the same way a sentence is split into words and even letters. The proposed method employs time-wise features extracted directly from the model we want to pre-train to predict the order of a shuffled action at the frame level (example in Fig. 1). Additionally, our approach can predict frames or chunks of the action from arbitrary parts of the video, which completely disregards any constraint related to allowing different factorization orders of the distribution. Fig. 2 illustrates the overview of the proposed framework applied to the vanilla spatiotemporal Graph Convolutional Networks (GCNs). Our work makes three main contributions. Firstly, we propose a novel and scalable atomic motion representation learning approach for 3D human action recognition. Secondly, we present a methodology that can be easily integrated and replicated into any skeleton-based model. Finally, we demonstrate the effectiveness of our approach through semi-supervised and self-supervised learning, achieving significant improvements in action recognition performance on two benchmark datasets: NTU RGB + D [37] and NTU-120 RGB + D [38]. Our results show that ATOM outperforms the current state-of-the-art methods in both datasets.

2. Related work

The use of self-supervised learning skeleton-based behaviour analysis intends to model 3D joint coordinates from unlabeled data. Current methods learn to extract structural and dynamic patterns by applying multiple video-based SSL pretext tasks and incremental modules into the base architecture they want to fine-tune to solve those tasks.

2.1. Skeleton-based behaviour analysis

The human body pose is one of the promising research lines in behaviour analysis. This kind of structured representation is semantically rich and very descriptive of human dynamics, attenuating appearance noises that RGB and depth data contain and, consequently, driving the learning process solely over human behaviour. Skeleton-based behaviour analysis rapidly evolved over the last decade from pseudo-images with CNNs [39–42] and sequence coordinate vectors with RNNs [43–47] to the solid improvements of GCNs [48,49,32,30,50–55], which learn to better represent embedded structural information by modeling skeleton data as a spatiotemporal graph.

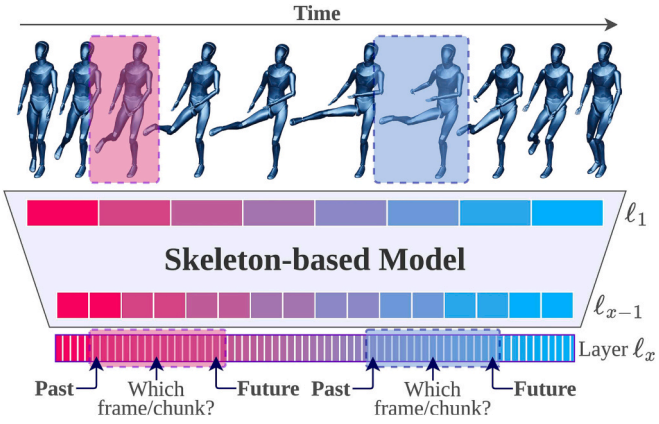

Fig. 1. Atomic order prediction. In a single forward pass of the whole action, we predict the order of its frames/chunks. The proposed ATOM scheme is able to split and learn ordering up to 64 frames/chunks, which is the key to obtain improved action representations.

However, self-supervised learning approaches still apply video-based techniques, where RGB (and depth) data is far more complex than positional joints information.

2.2. Self-supervised learning

Despite promising early results on self-supervised learning in NLP [10–13] and image [14–16] fields, such improvements in human behaviour analysis are yet to be achieved due to data complexity. Current image-based techniques solve jigsaw problems [21] to learn spatial relationships, image colourisation [56] by mapping objects to colours and further data transformations, such as scaling, inpainting and warping [57] through contrastive learning to construct positive and negative pairs. Some extensions of 2D techniques to 3-dimensional data [58,59] were also applied but lacked extracting temporal information. More recently, video-based SSL were proposed through playback speed [25], cycle consistency [60] and motion continuity [61] to pay more attention to the temporal axis. Since skeleton data is obtained from RGB and depth sources, naturally, our intuition leads us to apply SSL techniques from previous image and video-based works. However, as previously stated, such exploration towards self-supervised skeleton-based behaviour analysis is still minimal with concerning limitations. Previous works focused on motion prediction [34,27] with an autoencoder architecture to learn an encoder that is later fine-tuned to a supervised task. However, designing such a decoder lacks reproducibility for other encoder architectures and sometimes requires repeatedly conceiving a new decoder. Applying solely RGB video-based SSL techniques [29] (e.g., jigsaw puzzle), may become suboptimal and suffer from limitations [28] at both spatial and temporal levels (as shown in Table 3). Additionally, the increased number of parallel tasks will restrict the power of the pre-trained representations, especially for the fine-tuning approaches, becoming even more complex to solve [21]. This paper proposes a novel approach that explores fine-grained information at the frame level without additional architecture modules and computational complexity to easily pre-train and become greatly beneficial. The video-based works most related to ours are those which try to predict the time order [22–24]. Furthermore, we expect to offer a new direction for learning standards in a self-supervised paradigm for skeleton-based behaviour analysis.

3. Proposed method

SSL methods learn valuable representations when applied to non-trivial and non-ambiguous concepts [21], taking into account the type of data and problem. ATOM aims to improve the recognition accuracy of supervised methods while increasing the easiness of weight transferability into different models through atomic behaviour analysis and prediction. This section presents 1) the proposed approach, introducing the vanilla spatiotemporal Graph Convolutional Network (GCN) adopted, 2) the proposed atomic methodology, and 3) the improvements made in the ATOM module.

3.1. Graph convolutional networks preliminaries

GCNs are currently the state-of-the-art for 3D human action recognition, and our proposed ATOM module is not limited to this model type, but can also be applied to other model types such as pseudo-image-based, autoregressive, generative, and more. For consistency purposes, we adopt a notation as close as possible to previous GCN works [30–32]. A spatiotemporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denotes the skeleton data with N joints and T frames. Accordingly, the skeleton sequence's feature map is represented as $\mathbf{X} \in \mathbb{R}^{N \times T \times C}$, where C is the number of channels describing the joints coordinates. A GCN is comprised of both spatial and temporal graph convolutions, where an adjacency matrix $\mathbf{A} \in \{0,1\}^{N \times N}$ and corresponding identity matrix $\mathbf{I}$ determine the intra-body joints

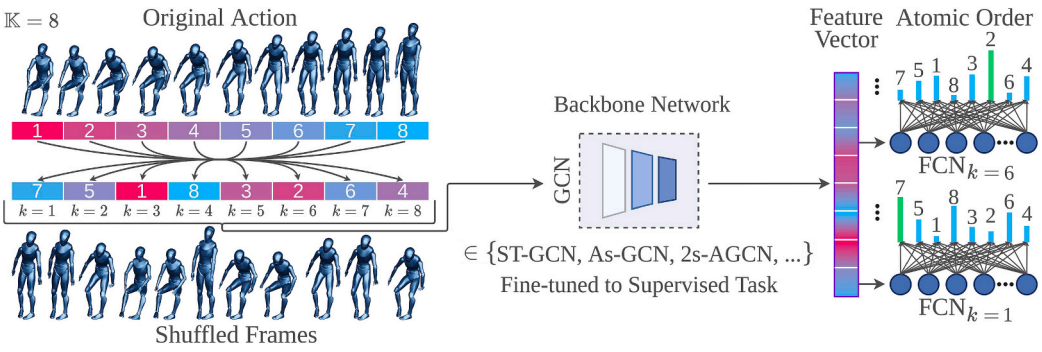

Fig. 2. Cohesive view of the proposed atomic methodology. Example of applying the proposed ATOM module to a skeleton-based model (e.g. ST-GCN [30], As-GCN [31] or 2s-AGCN [32]) with an action divided into $k = 8$ temporal chunks: given an action sample $\mathbf{X}$ with T frames, the sequence is shuffled into chunks of T/k framesor at the frame level, where $k = T$. The chunks are then concatenated and fed the model $\mathscr{M}$ we want to pre-train. The extracted time-wise feature vector is split into k sub-vectors and fed into k different standard Softmax layers, where each one predicts the corresponding temporal index k.

connections over the spatial dimension. For a single frame, the graph convolution is computed as:

$$\mathscr{S}(\mathbf{X}) = \sum_{i=1}^{p} \Lambda_i^{-\frac{1}{2}}(\mathbf{A}_i \odot \mathbf{M})\Lambda_i^{-\frac{1}{2}}\mathbf{X}\mathbf{W}_i, \tag{1}$$

where the degree matrix $\Lambda_p^{ii} = \sum_j(\mathbf{A}_p^{ij})$ normalises the adjacency matrix $\mathbf{A}_p$ through the number of edges attached to each joint node. $\mathbf{W}_p$ denotes the stacked weight vectors for each partition group p (spatial configuration proposed by [30]), and $\mathbf{M} \in \mathbb{R}^{N\times N}$ is a learnable weight matrix (initialised as an all-one matrix) on each layer.

Since consecutive frames describe consecutive skeletons, one-dimensional kernels are used as the temporal graph convolution, which is applied over the temporal axis after the spatial graph convolution. The spatiotemporal graph convolution is given by convolving the positional features joint-wise as:

$$\mathscr{T}(\mathscr{S}(\mathbf{X})) = \mathscr{S}(\mathbf{X}) * \mathbf{w}, \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^{1\times t\times C}$ is the temporal kernel with t as the number of frames to be convolved in the kernel. Eq. (2) will be further employed to extract the last layer of the corresponding model $\mathscr{M}$ we want to pre-train. The rationale is to obtain temporal features from spatial receptive fields and, therefore, dynamically operate at the atomic motion representation level directly from the model itself.

3.2. Atomic motion representation

Recent well-known self-supervised language representations [10,13] feed entire sequences directly to the model (as a bidirectional approach), achieving unprecedented results, such as in sentence order prediction. However, self-supervised video-based order prediction [22–24] approaches will directly divide the input into multiple clips or chunks before feeding it to a model and working in a multi-stream paradigm to solve the desired task. Despite being understandable in video data, due to its complexity and achieving a more granular level of information, feature concatenations are manually structured and will lack spatio-temporal dependencies between clips. ATOM consists of a self-supervised technique using atomic motion representation. As previously expressed, we assume skeleton-data closer to a text data concept rather than video data in pre-training representation learning. This hypothesis is supported by the fact that skeleton-based models define a fixed number of joints ($\mathscr{V}$) and edges ($\mathscr{E}$) across time for every human present in the data, which is far simpler than pixel data. Additionally, this is one of the reasons why some autoregressive approaches [47,62] in human action recognition still outperformed image-based algorithms [63,42] before GCNs became the mainstream. Since temporal graph convolutions are computed joint-wise over spatial graph convolutions, we compute Eq. (2) from the last layer of any model used in our experiments and, inherently, obtain a sequential temporal representation (as shown in Fig. 1). Such temporal feature representations extracted from a whole action $\mathbb{X} = \langle x_1, x_2, \ldots, x_n\rangle$ with n dimensions allow us to grasp and directly work with dynamic features x_n at the atomic level without compromising any temporal dependencies, instead of working with multiple forward passes of splitted action chunks and subsequently concatenated.

3.3. Atomic order prediction

The proposed atomic representation allows the model to learn on its own without manually structured features (e.g., pairwise feature concatenation [24,22]) or modifications to the current architecture. Since temporal features are obtained from spatial features, which are extracted directly from the skeleton data, it enables modelling bidirectional temporal dependencies from any segment of the action. For a given action sample $\mathbf{X}$ with T frames, the number of different orders $T!$ is factorial. However, instead of treating the order prediction task as an order classification, the task is formulated as an independent index classification problem (see Fig. 2), where the model can classify any index to any frame across the action. Previous works [22] claim that this proxy task at the frame level, and even by dividing the action into 4 clips, becomes too hard to solve with video data. Yet, since we are directly working on human dynamics (skeleton), the complexity is softened and becomes optimal to work at such a granular level. Specifically, for an action sequence $\mathbf{X}$, we define the input to a model $\mathscr{M}$ as the concatenation of shuffled frames or temporal chunks $\mathbf{S}$. When the order is predicted at the chunk level instead of frame level, we denote a temporal chunk k as $\mathbf{X}_k \in \mathbb{R}^{N\times k\times C} = \{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_k\}, k \in \mathscr{Z} \wedge \forall k \mathrm{mod} T = 0 \wedge 1 < k \geqslant T$. As illustrated in Fig. 2, the core idea of ATOM's methodology is to perform independent index classification, and as a result of our atomic motion representation, we are able to output a probability distribution for each frame or chunk. After extracting a temporal feature vector from model $\mathscr{M}$, we split it into k subvectors, which will implicitly correspond to the atomic representation of each $\mathbf{S}_k$ and inherently possess spatiotemporal features from previous and following chunks. Such bidirectional features can dynamically capture spatiotemporal dependencies, where we then use a linear layer k for each frame/chunk k as follows:

$$\boldsymbol{s}_k = \mathscr{W}_k \mathscr{M}(\mathbf{S})_k + \boldsymbol{b}_k, \tag{3}$$

where $\mathscr{W}_k$ and $\boldsymbol{b}_k$ are the parameters of linear transformation k. The probability distribution is then calculated to predict the index k being corrected as:

$$\boldsymbol{p}_k = \frac{\exp(\boldsymbol{s}_k)}{\sum_{l=1}^{\mathbb{K}} \exp(\boldsymbol{s}_l)}, \tag{4}$$

where $\mathbb{K} = \{1, 2, \ldots, k\}$ is the total number of frames/chunks to be ordered. Since the same model $\mathscr{M}$ parameters are shared across all linear

layers, and index orders during training, s_k has information on every other $s_l, l \neq k$ in the sequence, hence being able to capture bidirectional context. The assumption behind this is that the model $\mathscr{M}$ must jointly learn to understand the underlying spatial content of the skeleton alongside its motion across the shuffled action before they can distinguish between frames/chunks. Formally, we define independent index classification by the mean loss of each predicted index from each frame/chunk as:

$$\mathscr{L} = \frac{1}{\mathbb{K}} \sum_{l=1}^{\mathbb{K}} \left(- \sum_{k=1}^{\mathbb{K}} y_l^k \log(\boldsymbol{p}_k^l) \right), \tag{5}$$

where y_l^k and $\boldsymbol{p}_k^l$ are the probability of the action chunk k belonging to index of shuffled frame/chunk l in groundtruth and predction and $\mathbb{K}$ the number of total frames/chunks. The loss $\mathscr{L}$ is then backpropagated to optimise the whole ATOM framework. When the framework is trained to predict the order of such fine-grained information, the model $\mathscr{M}$ is trained to extract meaningful and precise spatiotemporal features and builds an approximation form of perception from human skeleton data at the frame level without action labels. Subsequently, since any modification was made to the skeleton-based architecture, the weights are transferred to fine-tune over a supervised task with apriori knowledge achieved.

4. Experiments and discussion

In this section, extensive experimental evaluations are performed to validate the proposed approach on four benchmarks. The ablation studies examine the contributions and efficacy over both benchmarks of NTU RGB + D [37]. Then, our best performing model is compared to the state-of-the-art approaches over NTU RGB + D [37] and NTU-120 RGB + D [38] under different settings.

4.1. Dataset, evaluation metrics and experimental settings

NTU RGB + D[37]. The dataset contains 56,880 video samples across 60 action classes, each with 3D skeleton data from 40 volunteers and 25 joints per skeleton. Two recommended benchmarks are provided: 1) cross-subject, which trains models on 20 subjects and tests on the remaining ones, and 2) cross-view, which trains on camera views 2 and 3 and tests on view 1.

NTU-120 RGB + D[38]. This dataset is an extended version of its predecessor, comprising 114,480 video samples with 120 action classes, captured with three camera views in 32 different setups and 106 volunteers with 25 body joints for each skeleton. The dataset proposes two benchmarks: 1) cross-subject, where models are trained on 53 subjects and tested on the remaining ones, and 2) cross-setup, where models are trained on even setup IDs and tested on odd setup IDs.

The datasets are used in their original format, which is already presented in a user-friendly and compact form that can be accessed at https://rose1.ntu.edu.sg/dataset/actionRecognition/.

Evaluation metrics. In accordance with most skeleton-based action recognition approaches, we adopt the Top-1 and Top-5 accuracy for metrics for evaluating the recognition performance. The conventional Top-1 accuracy assesses the order prediction. **Experimental settings**. Two different settings are used to validate the proposed framework. (1) Self-supervised pre-training: the backbone network is initialised with the learned weights from the self-supervised task instead of training from scratch and randomly initialising the network's weights. We then learn the network for action recognition. (2) Semi-supervised learning: The network is pre-trained without any action labels and using the learned weights to initialise the classifier, which will be trained on small percentages of training data (5% and 10%).

Implementation details. Our framework is implemented with PyTorch [64] to facilitate the computation comparisons and transferability between other backbone networks, which were also developed with the same library. The proposed method for order prediction uses 64 frames as input since its the action temporal execution average in NTU RGB + D [37] and NTU-120 RGB + D[38], where the maximum length is 300 frames. For this reason, normalising to 0 the remaining frames or cycling those actions will naturally disturb the learning process of the order prediction task. We evaluate the proposed framework with three backbone networks: ST-GCN [30], As-GCN [31] and 2s-AGCN [32]. It is worth mentioning that modifying the optimisation strategy, hyperparameters or learning rate schedulers during pre-training may lead the trained weights to local minimums when transferring to the supervised task. Therefore, we adhere strictly to the original training and evaluation settings of each architecture as specified in the respective dataset. Apart from the ablation studies, all reported results in this paper use $k = 8$ temporal chunks with configuration (D) from Table 1 if not mentioned otherwise.

4.2. Ablation study

Before diving into state-of-the-art comparisons, we first demonstrate experimentally that our proposed ATOM's properties considerably improve action recognition.

Different framework designs. In Table 1, we compare the action recognition accuracy for various framework architectures in both benchmarks of NTU RGB + D [37] under self-supervised pre-training settings. The "supervised" baseline (ST-GCN [30]) represents the backbone network trained from scratch and randomly initialising the baseline's weights. We start with our proposed configuration (A) by applying action sequence order prediction to skeleton data with divided action chunks under the order classification paradigm. The backbone network is fed with one clip each at a time, and the extracted features are pairwise concatenated (similar to [24,22]). Since our goal is to achieve atomic motion representation by using a more significant number of chunks than [24,22], we then confirm the slightly improved performance by concatenating those features in the correspondence shuffled order before feeding to a single FCN (B) to perform order classification. After observing the increased flexibility (over skeleton data) of the framework dynamically operating through the concatenated features, we also concatenated shuffled chunks and extracted a feature vector with a single forward pass (C). Finally, it leads us to introduce single order index classification through an adaptive FCN, where each frame or

Table 1
Evaluating different architecture designs on NTU RGB + D [37]. Except for the "supervised" baseline, each configuration is pre-trained with order prediction before being applied to the action recognition task.

	Arch	Method	*Top* −1	*Top* −5	*Top* −1	*Top* −5
NTU RGB-D			*Cross-Subject*		*Cross-View*	
	ST-GCN	Supervised [30]	81.51	96.92	88.34	98.33
		A Proposed	82.36	96.99	88.75	98.61
		B + Single FCN	82.53	97.04	89.07	98.72
		C + Adaptive Feat.	82.93	97.09	89.58	98.81
		D + Adaptive FCN	**83.24**	**97.13**	**90.36**	**99.08**

chunk has a corresponding linear layer (D) to classify such fine-grained information and distinctively overcome the baseline.

Order prediction effectiveness. After achieving such improvements in skeleton-based order prediction, we also test the limits of our framework by increasing the number of indexes to be predicted across both benchmarks of NTU RGB + D [37]. Naturally, even for humans, the difficulty of the problem proportionally increases. Considering the input to our network is an action sequence of 64 frames, we perform order prediction through the power of 2 action tuples, as illustrated in Fig. 3. The proposed ATOM framework can almost perfectly predict the order of up to 8 different action chunks. Considering that the accuracy of random guessing for the task of 16 different chunks is 6.3%, the framework notably learns precise information and achieves 75.6%. Despite the drop in accuracy with 32 splits and at the frame level (64 frames), we can reach 18.9% and 9%, respectively, showing that ATOM indeed learns to analyse the increasing fine-grained information, which can also be positively regarded when taking into account the accuracy of random guessing.

Additionally, we also verify how effective can the proposed framework be in proportion to the increased number of chunks used to pre-train the backbone network. As shown in Fig. 4, even with the most straightforward task (action split in 2) and the most complex task (frame-level, 64 frames), the proposed strategy is still capable to boost the action recognition performance. The peak performance was achieved when pre-training the backbone network to predict 8 shuffled chunks. This indeed confirms the difference in self-supervised learning modalities between skeleton-based data and video data, where previous work [22] achieved better performance with only 3 shuffled clips in a total of 48 frames (16 frames each clip).

Skeleton structure learning. Finally, we also analyse directly how a backbone network performed over a pretext task, such as order prediction, can learn the importance of physical connections of the human body compared with the action recognition task, where intuitively, we provide much more information about human dynamics. Fig. 5 shows an example of the adjacency matrices learned for the cross-subject benchmark of NTU RGB + D [37]. The green scale of each element in the matrix represents the strength of the connection between respective joints. The left matrix is the original adjacency matrix from the baseline [30] trained for action recognition from scratch. The middle matrix is our framework (using the baseline as the backbone network) trained for order prediction. The right matrix is our baseline initialised with the pre-trained weights for action recognition. It is clear that the learned structure of the graph with solely order prediction (middle) is a bit more flexible (lighter colour) while containing a similar pattern as the baseline trained from scratch (left) for action recognition despite far less information provided about the skeleton. Furthermore, the weight transferability (right) for action recognition can also be regarded as beneficial since we can distinctly observe the strengthening between adjacent joints.

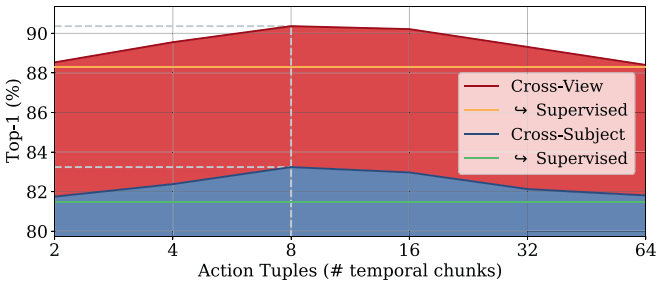

Fig. 4. Performance evolution increasing action chunks on NTU RGB + D [37]. The horizontal lines correspond to the respective performances without self-supervision on each benchmark. All results apply $k = 8$ temporal chunks, if not mentioned otherwise.

4.3. Comparison with the state-of-the-art

We perform the evaluation of our best model architecture effectiveness with respect to the previous state-of-the-art approaches of supervised, self-supervised and semi-supervised learning towards skeleton-based action recognition over four evaluation protocols from NTU RGB + D [37] and NTU-120 RGB + D[38].

4.3.1. Self-supervised learning

As previously described, we evaluated our method termed **ATOM** combined with three GCN-based architectures and directly compared them with the latest SSL techniques in the video and skeleton domain. As shown in Table 2, our ATOM methodology achieves the best results on every backbone network over the two benchmarks of NTU RGB + D [37]. The first observation was the slight decrease and increase in performance from pace prediction techniques [25] over skeleton data, confirming our claims that applying video-based techniques to the human skeleton does not always work. Furthermore, the superiority of our model compared with multi-task methods (MCC) [27], shows the efficacy of avoiding solving multiple problems at once, which can lead to an ambiguous concept [21] over the self-supervised pre-training process. In our view, this is mostly due to the fact that they adopted a pace prediction module in conjunction with a motion continuity module to learn a decoder to reconstruct the interpolated motion, which can disturb the learning process of the decoder and, consequently, the encoder. Additionally, it is also challenging to generalise a decoder that suits multiple backbone networks at once, increasing the difficulty of designing new decoders, e.g., multi-stream architectures (2s-AGCN [32]).

Table 3 shows the results obtained in a far more challenging dataset

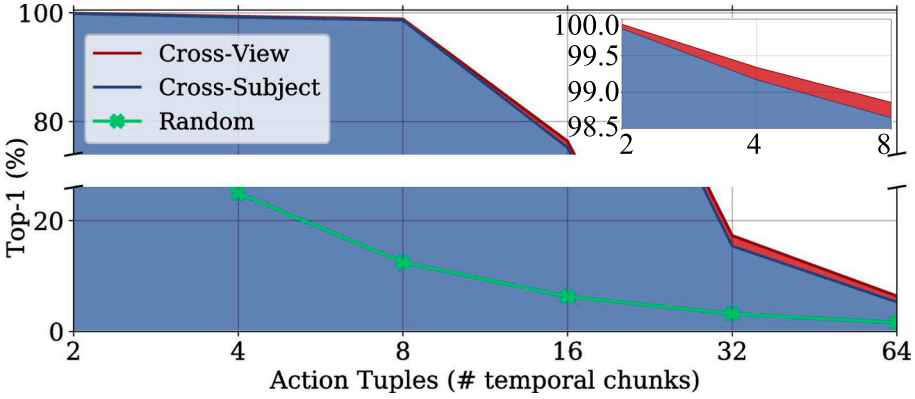

Fig. 3. Order prediction accuracy increasing action chunks on NTU RGB + D [37]. The green line correspond to the respective accuracy of random guessing.

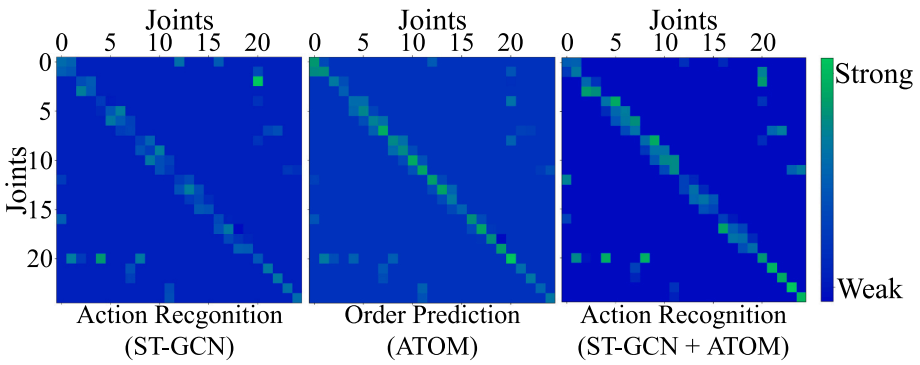

Fig. 5. Learned weight importance adjacency matrix on the cross-subject benchmark of NTU RGB + D [37]. The strength between connected joints is represented by the green scale. The left matrix is the original learned adjacency matrix. The middle matrix is ATOM (ST-GCN as backbone) trained on the pretext task. The right matrix is the corresponding matrix learned by initialising the weights with ATOM for action recognition.

Table 2
Action recognition performance comparison on NTU RGB + D[37]. The first row of each backbone network represents the results trained from scratch.

	Arch	Method	Top −1	Top −5	Top −1	Top −5
NTU RGB-D			*Cross-Subject*		*Cross-View*	
	ST-GCN	Supervised [30]	81.51	96.92	88.34	98.33
		VCOP [22]	82.24	97.12	88.61	98.54
		Jigsaw Puzzle [65]	81.78	96.98	89.02	98.57
		Pace Prediction [25]	81.54	96.97	88.77	98.55
		MCC [27]	83.01	97.05	89.68	98.84
		ATOM (ours)	**83.24**	**97.13**	**90.36**	**99.08**
	As-GCN	Supervised [31]	86.76	97.58	94.15	98.99
		VCOP [22]	87.51	98.14	94.91	99.03
		Jigsaw Puzzle [65]	87.06	97.96	94.61	99.02
		Pace Prediction [25]	87.31	98.03	95.01	99.13
		MCC [27]	88.37	98.39	95.45	99.26
		ATOM (ours)	**88.53**	**98.45**	**95.68**	**99.39**
	2s-AGCN	Supervised [32]	88.51	99.43	95.12	99.14
		VCOP [22]	88.98	98.73	95.76	99.43
		Jigsaw Puzzle [65]	88.81	98.60	95.35	99.31
		Pace Prediction [25]	89.18	98.64	95.55	99.30
		MCC [27]	89.66	98.80	96.27	99.45
		ATOM (ours)	**89.79**	**98.90**	**96.46**	**99.56**

Table 3
Action recognition performance comparison on NTU-120 RGB + D[38]. The first row of each backbone network represents the results trained from scratch.

	Arch	Method	Top −1	Top −5	Top −1	Top −5
NTU-120 RGB-D			*Cross-Subject*		*Cross-Setup*	
	ST-GCN	Supervised [30]	75.08	89.92	76.12	92.04
		VCOP [22]	76.04	92.15	76.82	93.73
		Jigsaw Puzzle [65]	76.26	92.45	77.07	94.99
		Pace Prediction [25]	75.78	90.88	75.89	91.24
		MCC [27]	77.02	94.93	77.78	95.83
		ATOM (ours)	**77.65**	**95.64**	**79.44**	**96.52**
	As-GCN	Supervised [31]	78.37	95.64	78.92	96.12
		VCOP [22]	78.37	98.14	80.01	97.02
		Jigsaw Puzzle [65]	78.55	99.29	79.87	96.98
		Pace Prediction [25]	78.04	96.11	79.75	96.83
		MCC [27]	79.37	96.42	80.81	97.23
		ATOM (ours)	**79.98**	**96.96**	**81.45**	**97.96**
	2s-AGCN	Supervised [32]	79.55	96.62	81.73	96.91
		VCOP [22]	80.57	97.13	82.49	97.93
		Jigsaw Puzzle [65]	80.78	97.21	82.36	97.83
		Pace Prediction [25]	80.29	97.10	82.08	97.68
		MCC [27]	81.25	96.65	83.26	97.43
		ATOM (ours)	**82.01**	**97.73**	**84.08**	**98.07**

Table 4
Semi-supervised learning evaluation on NTU RGB + D [37] and NTU-120 RGB + D [38]. Top-1 accuracy reported from small percentages of training data. The first row of each backbone network represents the results trained from scratch.

			Labelled Data			
	Arch	Method	5%	10%	5%	10%
NTU RGB-D			*Cross-Subject*		*Cross-View*	
	ST-GCN	Supervised [30]	38.19	52.41	40.42	56.94
		MCC [27]	42.40	55.57	44.71	59.85
		ATOM (ours)	**42.96**	**56.33**	**45.12**	**60.13**
	As-GCN	Supervised [31]	41.11	55.74	44.73	59.49
		MCC [27]	45.47	59.18	49.15	63.06
		ATOM (ours)	**46.15**	**60.46**	**50.02**	**64.16**
	2s-AGCN	Supervised [32]	43.51	57.24	49.07	62.03
		MCC [27]	47.39	60.75	53.31	65.79
		ATOM (ours)	**47.92**	**61.33**	**54.11**	**66.27**
NTU-120 RGB-D			*Cross-Subject*		*Cross-Setup*	
	ST-GCN	Supervised [30]	38.19	52.41	40.42	56.94
		MCC [27]	42.40	55.57	44.71	59.85
		ATOM (ours)	**43.74**	**57.12**	**46.23**	**62.13**
	As-GCN	Supervised [31]	41.11	55.74	44.73	59.49
		MCC [27]	45.47	59.18	49.15	63.06
		ATOM (ours)	**46.93**	**61.19**	**51.03**	**64.95**
	2s-AGCN	Supervised [32]	43.51	57.24	49.07	62.03
		MCC [27]	47.39	60.75	53.31	65.79
		ATOM (ours)	**48.45**	**62.13**	**54.82**	**67.23**

(NTU-120 RGB + D [38]) with 120 different classes. Once again, the results obtained show the superiority of our framework on every backbone over both benchmarks. Despite similar results on order prediction, the ATOM method reaches performance increases up to 4.4% (3.3% p. p.) over the supervised baseline. We justify this phenomenon with the increasing amount of training data compared to NTU RGB + D [37], which is also explored in the next section. Overall, our experiments point out that using single tasks specifically designed for skeleton-based data employed for pre-training becomes more promising than applying multiple tasks from video-based techniques.

4.3.2. *Semi-supervised learning*

Since, in most situations of real-world applications, labelled data is not always available, training data-driven architectures becomes rather problematic. The evaluation in such conditions is a crucial experiment nowadays and not much adopted. For this reason, we also explore the proposed framework applied to small amounts of labelled data, specifically with 5% and 10%, to train our backbone networks. Table 4 reports the Top-1 accuracy results obtained under semi-supervised settings. The first conclusion is the poor performance of the baselines trained from scratch when data is insufficient. After transferring the pre-trained weights to the backbone networks, our framework is able to boost the recognition performance up to 14.1% (6.3% p.p.). From our perspective, the superiority of ATOM over previous SSL techniques is due to the fact that the reconstruction module of [27] is not so efficient when learning to generate human actions in such small percentages of data when trying to apply handcrafted features such as motion consistency and continuity. Consequently, learning an encoder in such conditions becomes suboptimal when compared to ATOM's single task technique.

5. Conclusions and further work

Whereas the recognition and prediction of human actions have been the focus of significant research efforts in the security/surveillance domains, it is still required a substantially large number of training samples. This paper introduced a novel atomic motion representation approach for self-supervised human action recognition. The rationale is to dynamically leverage fine-grained information from the atomic representation extracted by the backbone network itself, where we are able to better model skeleton data without any labels, obtaining improved representations of human actions. As a result, the proposed methodology becomes more suitable for skeleton data even with a single pretext task. Also, ATOM scheme is able to overcome previous limitations by facilitating weight transferability across multiple different architectures without any modification to the proposed framework. Our method was evaluated on four well-known benchmarks from NTU RGB + D [37] and NTU-120 RGB + D[38], significantly advancing the state-of-the-art performance metrics. In order to extend the applicability of the proposed ATOM approach, future research could focus on analysing the performance of the method on longer action sequences beyond the current limit of 300 frames. This could provide insights into the scalability of the self-supervised task and further enhance the effectiveness of the approach in real-world scenarios. Furthermore, to enhance the performance of existing SSL methods, it would be beneficial to investigate the impact of the quality of the backbone network used for feature extraction, and to explore the use of alternative modalities or active learning strategies for selecting the most informative samples.

CRediT authorship contribution statement

Bruno Degardin: Conceptualization, Methodology, Formal-analysis, Data-curation, Writing-original-draft. **Vasco Lopes:** Validation, Methodology, Writing-review-editing. **Hugo Proença:** Writing-review-editing, Funding-acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by the FCT/MEC through National Funds and by the FEDER-PT2020 Partnership Agreement under the Project CENTRO-01–0247-FEDER-113023 - DeepNeuronic. This work was also supported by FCT/MCTES through national funds and co-funded by EU funds under the project UIDB/50008/2020. This research was also supported by 'FCT - Fundação para a Ciência e Tecnologia' through the research grant 'UI/BD/150765/2020' and '2020.04588.BD'.

References

[1] D. Tran, H. Wang, L. Torresani, M. Feiszli, Video classification with channel-separated convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5552–5561.
[2] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
[4] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6202–6211.
[5] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark, arXiv preprint arXiv: 1609.08675.
[6] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv: 1705.06950.
[7] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, arXiv preprint arXiv: 1808.01340.
[8] J. Carreira, E. Noland, C. Hillier, A. Zisserman, A short note on the kinetics-700 human action dataset, arXiv preprint arXiv: 1907.06987.
[9] C. Gu, C. Sun, D.A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al., Ava: A video dataset of spatio-temporally localized atomic visual actions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6047–6056.
[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805.
[11] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv: 1909.11942.
[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv: 1907.11692.
[13] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Adv. Neural Inf. Process. Syst. 32 (2019).
[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[15] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.
[16] H. Bao, L. Dong, F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv: 2106.08254.
[17] Q. Kong, W. Wei, Z. Deng, T. Yoshinaga, T. Murakami, Cycle-contrast for self-supervised video representation learning, Adv. Neural Inf. Process. Syst. 33 (2020) 8089–8100.
[18] Y. Lin, X. Guo, Y. Lu, Self-supervised video representation learning with meta-contrastive network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8239–8249.
[19] Z. Liang, M. Yin, J. Gao, Y. He, W. Huang, View knowledge transfer network for multi-view action recognition, Image Vis. Comput. 118 (2022), 104357.
[20] N. Behrmann, J. Gall, M. Noroozi, Unsupervised video representation learning by bidirectional feature prediction, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1670–1679.
[21] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: European conference on computer vision, Springer, 2016, pp. 69–84.
[22] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, Y. Zhuang, Self-supervised spatiotemporal learning via video clip order prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10334–10343.
[23] K. Hu, J. Shao, Y. Liu, B. Raj, M. Savvides, Z. Shen, Contrast and order representations for video self-supervised learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7939–7949.
[24] H.-Y. Lee, J.-B. Huang, M. Singh, M.-H. Yang, Unsupervised representation learning by sorting sequences, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 667–676.
[25] S. Benaim, A. Ephrat, O. Lang, I. Mosseri, W.T. Freeman, M. Rubinstein, M. Irani, T. Dekel, Speednet: Learning the speediness in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9922–9931.
[26] A. Ben Tanfous, A. Zerroug, D. Linsley, T. Serre, How and what to learn: Taxonomizing self-supervised learning for 3d action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2696–2705.
[27] Y. Su, G. Lin, Q. Wu, Self-supervised 3d skeleton action representation learning with motion consistency and continuity, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13328–13338.
[28] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, J. Feng, Adversarial self-supervised learning for semi-supervised 3d action recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 35–51.
[29] L. Lin, S. Song, W. Yang, J. Liu, Ms2l: Multi-task self-supervised learning for skeleton based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2490–2498.
[30] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.
[31] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.
[32] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12026–12035.
[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67.
[34] K. Su, X. Liu, E. Shlizerman, Predict & cluster: Unsupervised skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9631–9640.
[35] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: A survey, Image Vis. Comput. 60 (2017) 4–21.
[36] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.
[37] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.
[38] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2684–2701.
[39] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3288–3297.
[40] C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 786–792.
[41] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recogn. 68 (2017) 346–362.
[42] T. Soo Kim, A. Reiter, InterpreTable 3d human action analysis with temporal convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 20–28.
[43] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1110–1118.
[44] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: European conference on computer vision, Springer, 2016, pp. 816–833.
[45] J. Liu, G. Wang, P. Hu, L.-Y. Duan, A.C. Kot, Global context-aware attention lstm networks for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647–1656.
[46] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 31, 2017.
[47] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive recurrent neural networks for high performance human action recognition from skeleton data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2117–2126.
[48] Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1113–1122.
[49] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183–192.

[50] B. Degardin, J. Neves, V. Lopes, J. Brito, E. Yaghoubi, H. Proença, Generative adversarial graph convolutional networks for human action synthesis, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1150–1159.
[51] H.-G. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, K. Ramani, Infogcn: Representation learning for human skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20186–20196.
[52] X. Zhang, C. Xu, D. Tao, Context aware graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14333–14342.
[53] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1112–1121.
[54] N. Sun, L. Leng, J. Liu, G. Han, Multi-stream slowfast graph convolutional networks for skeleton-based action recognition, Image Vis. Comput. 109 (2021), 104141.
[55] B. Degardin, V. Lopes, H. Proença, Regina—reasoning graph convolutional networks in human action recognition, IEEE Trans. Inf. Forensics Secur. 16 (2021) 5442–5451.
[56] G. Larsson, M. Maire, G. Shakhnarovich, Learning representations for automatic colorization, in: European conference on computer vision, Springer, 2016, pp. 577–593.
[57] S. Jenni, H. Jin, P. Favaro, Steering self-supervised feature learning beyond local pixel statistics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6408–6417.
[58] J. Walker, A. Gupta, M. Hebert, Dense optical flow prediction from a static image, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2443–2451.
[59] D. Pathak, R. Girshick, P. Dollár, T. Darrell, B. Hariharan, Learning features by watching objects move, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2701–2710.
[60] X. Wang, A. Jabri, A.A. Efros, Learning correspondence from the cycle-consistency of time, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2566–2576.
[61] H. Liang, N. Quader, Z. Chi, L. Chen, P. Dai, J. Lu, Y. Wang, Self-supervised spatiotemporal representation learning by exploiting video continuity, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 1564–1573.
[62] I. Lee, D. Kim, S. Kang, S. Lee, Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1012–1020.
[63] H. Liu, J. Tu, M. Liu, Two-stream 3d convolutional neural network for skeleton-based action recognition, arXiv preprint arXiv: 1705.08106.
[64] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch.
[65] D. Kim, D. Cho, I.S. Kweon, Self-supervised video representation learning with space-time cubic puzzles, in: Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, pp. 8545–8552.